community access to their search program and for answering questions regarding the FASTA algorithm.

3.1.1.6 New User Manual

A new *User Manual* was produced this year and distributed free of charge to all current BIONET users at the end of June 1988. Unlike the old *BIONET Training Manual* which was organized by program, the *User Manual* is organized around scientific tasks and contains examples which demonstrate the integration of several programs on BIONET to accomplish the specified tasks. The following sections are included in the manual:

- Introduction
- Mechanics (Files Structure and Operating System topics)
- Sequence Entry and Editing
- Sequence Location
- Sequence and Map Display
- Sequencing Project Management
- Sequence Translation
- Sequence Composition
- Sequence Structure
- Restriction Analysis
- Cloning Simulation
- Sequence Comparison

The new manual has been favorably received by the community as demonstrated by many positive comments received from the users by the BIONET consulting staff.

3.1.1.7 Revision of the*Introduction to BIONET*

In June, a completely revised *Introduction to BIONET* manual was sent to all subscribers along with the new *User Manual*. While the *User Manual* covers the use of scientific programs on BIONET, the *Introduction to BIONET* details aspects of accessing the system and the use of other BIONET features such as electronic communications. The *Introduction to BIONET* is given to each new lab group when they sign up for BIONET. At 100 pages in length, the new "Intro" is greatly expanded from its 70 page predecessor. A sample login session is included to assist the first time BIONET user, and a trouble-shooting chapter covers many of the common problems that might be encountered while using BIONET. Other new sections include information on contributed software and databases, utility programs, network electronic mail, and electronic submission of sequences to the national databases using the XGENPUB program. A new version of the *Introduction to BIONET* will be produced when BIONET ports over to the SUN computers which run the UNIX operating system.

3.1.1.8 Additional On-line HELP EXAMPLES

As part of the consultant's on-going effort to expand and improve the on-line documentation for BIONET, several additions have been made to the HELP EXAMPLES system of on-line examples of using various programs. The most important additions have been examples of the FASTA-MAIL program, the QUEST program, and the use of the FIND and RETRIEVE commands to locate and make personal copies of entries in the databases. The FASTA-MAIL example (on-line example no. 35) guides the user through submitting a FASTA-MAIL similarity search to the DNA or protein databases, and then accessing the search results through the MM mail program. Much of the explanation involves a discussion of the FASTA-MAIL scoring and alignment output and how to interpret the results. The QUEST example (example no. 33) takes the user through using a promoter "key" (consensus pattern) in the QUEST program to search the chimpanzee beta-globin gene nucleic acid sequence for polymeraseII promoter sites. Instructions are given for using the GUIDE command to set up the search and open a file in which to collect matched sequences, loading the promoter key pattern from IntelliGenetics' KeyBank database of consensus sequences, and running the search. The FIND and RETRIEVE example (example no. 40) takes the user through a typical sequence of locating a sequence of interest in the databases and then making a personal copy of the sequence that can be freely modified by the user. Although most database retrieval tasks can now be accomplished through the new Intelligenetics' FINDSEQ program (see below), the FIND command is still an extremely fast method of doing keyword searches, in addition to being useful for locating sequences in databases that are not yet handled by FINDSEQ such as KEYBANK and VECTORBANK. Constant revision and expansion of the on-line help system is needed as new programs and databases are added to the BIONET system.

3.1.1.9 BIONET Newsletter

The BIONET staff began producing a hardcopy newsletter called *BIONET News*. Two issues have been mailed to the user community (April and October 1988) and copies are included in *Appendix III.*

3.1.1.10 IDEAS programs

Five protein structure prediction programs from the IDEAS (Integrated Database and Extended Analysis System) suite written by Dr. Minoru Kanehisa of the National Cancer Institute and Kyoto University were released on BIONET in February. Since the IDEAS suite only runs on VAX/VMS systems, BIONET provided remote access to a networked MicroVAX owned by IntelliGenetics on which the suite is installed. To access the programs, BIONET users run a simple program which logs them into the MicroVAX. Data files may be moved to the MicroVAX and output results files can be transmitted back to the BIONET DEC-2065 for further analysis. Since its introduction, the IDEAS suite has been used an average of 23 times per month. Further details are provided below under **PC and Minicomputer-based Software.**

3.1.2 Collaborative Research

BIONET's collaborative community is made up of several components, encompassing efforts by outside scientists working in conjunction with BIONET staff. In subsequent sections we discuss each component in more detail:

- **Collaborative Efforts by the BIONET research staff.** This covers research performed by Drs. Jurka and Maulik together with outside investigators.
- **DEC-2065 Software Contributors.** This component includes those persons who have

contributed software for use by the BIONET community on the central DEC-2065 computer;

- **PC and Minicomputer-based Software**. This component includes our efforts to gather and disseminate software of special utility to the community;
- **Data Contributors**. This component includes those persons who, together with the BIONET staff, contribute data useful to the community;
- **Liaison with Other Resources**. Several accounts have been established to promote sharing of information among molecular biology computing resources;
- **Bulletin Boards and Leaders**. This component includes those persons who have agreed to maintain bulletin boards of special interest to scientists using BIONET.

3.1.2.1 Collaborative Efforts by the BIONET research staff

The BIONET research group which is headed by Dr. Jurka and includes Dr. Maulik and Mr. Liang Jen Horng, our Applications Programmer, has engaged in collaborative research on the evolution of Alu and protein sequences. Other preliminary investigations in the area of protein structure prediction have been pursued as well.

Dr. Jurka has developed collaborative research projects with Professor Roy Britten (CalTech, abstract by Jurka and Britten at Cold Spring Harb. Symp. May, 1988), and Dr. Emile Zuckerkandl (Linus Pauling Institute, abstract by Jerzy Jurka with Emile Zuckerkandl and Jerry Latter has been presented on an International FEBS Meeting, held in Sept. 1988 in Corsica). Collaborative research with both institutions is in progress and the abstracts describing this work are provided in *Appendix I*. Dr. Jurka has also been appointed to the editorial board of the *Journal of Molecular Evolution*. He has also reviewed manuscripts for *Genomics* and *Nucleic Acids Research*.

In collaboration with Aleksandar Milosavljevic and Professor David Haussler from the University of California at Santa Cruz, Dr. Jurka and Mr. Horng worked on applications of machine learning algorithms to the classification of biological sequences. In particular, they investigated a class of models of unsupervised learning, based on the principle of cognitive economy. Using prototype algorithms based on these models they were able to successfuly reproduce classification of aligned human Alu sequences (Jurka and Smith, 1988). They intend to explore the applicability of these models to the classification of unaligned sequences.

As part of an on-going research project designed to aid molecular biologists in predicting the structural and functional properties of biological molecules from their sequences, BIONET has created a "higher-order" database of protein structural characteristics. Derived from the Brookhaven database (PDB) of protein structures, the DSSP database is created using the algorithm of Kabsch and Sander to create a dictionary of secondary structure of those proteins that are known to be non-homologous at the level of sequence and structure. The dictionary describes the secondary structural patterns in terms of helix, beta-bridge, beta-ladder, turn, bend, and coil and is also annotated with comments taken directly from the principal literature citations describing the structure. These include general features of the protein as well as specific annotations of residues involved in active sites or regions, forming the core, etc. The annotations are in the format: predicate-name/argument(s)/terminator to make them easily readable by a suitable computer program. BIONET has begun preliminary investigations with the Research Institute for Advanced Computer Science (RIACS) at NASA/Ames to design a Lisp-based software package that will be able to infer structural features from this database.

3.1.2.2 DEC-2065 Software Contributors

The following are the major contributors of software for use by the community on the DEC-2065:

FASTP, FASTN, and FASTA From Bill Pearson - BIONET users have largely migrated from the older FASTP and FASTN programs this year to the FASTA-MAIL program which was described above in the **Service** section. The user interface for FASTA-MAIL runs on the DEC 2065 but the actual database searches are run remotely on BIONET's Sun 3/280 computer. The results are returned to the user on the DEC by electronic mail. The FASTA-MAIL program is used about 950 times per month or about 32 times per day!!

BIONET users are currently using FASTP on the DEC at the rate of 266 times per month. FASTN was removed from the DEC due to the advent of FASTA-MAIL and the wish to remove compute-intensive searches from the DEC but the FASTP program (which consumes far fewer resources and can be used to search user-created databases) was maintained.

MULTAN - MULTAN is a program developed by Dr. Bill Bains for aligning multiple homologous DNA sequences. While it is limited to being able to align sequences which are at least 60% homologous, it is an extremely rapid program. Multiple sequence alignment is useful for BIONET users studying evolution and for those trying to obtain a consensus from many sequences of similar function. A new version of MULTAN has been received from Dr. Bains and will be made available when BIONET users are transferred to the Sun computers.

BIOFOLD - Three years ago Dr. Michael Zuker, from the NRC laboratory in Ottawa Ontario made BIOFOLD available as a program on the BIONET computer. This program predicts RNA secondary structure and is used an average of 7 times per month.

ALIGN - Dr. Dan Davison, while a graduate student at Stony Brook at SUNY wrote and contributed two versions of his ALIGN program. The first version runs directly on the BIONET computer and can be used to align two very long DNA sequences including sequences which are not very homologous to each other and contain large gaps. The alignments are significantly better than similar heuristic alignments obtained from the SEQ SEARCH procedure in the BIONET core programs. The alignments are not as good as obtained from the SEQ ALIGN procedure but Dr. Davison's ALIGN program is significantly faster. The ALIGN program is used an average of 31 times per month.

XPROF - Dr. Rose, at Pennsylvania State University, has contributed the DEC-VAX Fortran version of his method for calculating hydropathicity profiles for proteins based on empirical observations on the extent to which amino acid residues are found to be exposed or buried. This program was released last year on the DEC and is used an average of 6 times per month.

XALIGN - As part of BIONET's effort last year to make available multiple sequence alignment software, we prepared and released XALIGN, based on Bacon and Anderson's ALIGN program. This program, as described in Bacon, D.J. and W.F. Anderson, J. Mol. Biol. 191: 153-161, 1986, finds the best local alignments among up to five protein sequences. The program was developed to run on most Fortran77-compatible computers and has been significantly modified to run on BIONET's DEC

2065 computer. XALIGN uses a large amount of CPU time and we thus asked the BIONET Community to use it during offpeak hours only, outside of 8 - 5 PST. The program has been used an average of 6 times per month since its release.

3.1.2.3 PC and Minicomputer-based Software

PC and minicomputer-based software is available to BIONET users via the BIONET lending library. This software runs on local computers instead of the DEC 2065 and, as such, helps reduce the demands on our central resource. Other minicomputer software (IDEAS, see below) is available to BIONET users on an IntelliGenetics-owned MicroVAX.

We began the lending library concept by making Kermit (a terminal emulator and file transfer protocol from Columbia University) available to BIONET users. BIONET copies Kermit onto diskettes provided by the users and returns the disks by regular mail. (Smaller lending library programs are also available for downloading from the DEC 2065.) We currently make Kermit available for Apple II, Macintosh, IBM PC and TRS-80 model computers. We have extended our lending library to include many BIONET user developed programs and a number of utility programs that are useful for file transfer on IBM PC and Macintosh computers. Software for VAX and Sun minicomputers has also been added to the lending library. A catalog of lending library software is available in *Appendix IV*. We report below the new contributed software during this past year.

MS-Kermit version 2.30

In May, version 2.30 of the public domain MS-Kermit communications package for IBM compatible computers was added to the BIONET Lending Library for free distribution to users who send in a blank diskette. This version of Kermit is significantly improved from its predecessor in that it is able to emulate a Tektronix 4010 graphics terminal. It also allows dialup procedures to be automated in a command file. The ability to emulate a graphics terminal means that BIONET users that have the Kermit package and an IBM PC or compatible can obtain the graphics output from programs such as DDMATRIX, CLONER, HPLOT, and HCOMP over their telephone connection. If they also have a graphics printing package, they can then print these graphic displays on their local dot matrix or laser printer. BIONET took advantage of the new macro capability in MS-Kermit to create login command files that allow users to dial a local Telenet or Compuserve number, and go through the complete BIONET login procedure by simply typing a single command to the Kermit program prompt. These login command files are given to users if they obtain Kermit from BIONET. Instructions for using the package with BIONET, and the original Kermit manual are also distributed to users. Since its release 74 requests for the new version of Kermit have been handled by the BIONET administrator.

New version of MacDNA

MacDNA is a program written by Robert Schleif from the Dept. of Biochemistry, Brandeis University. An updated version of MacDNA has been made available through the BIONET lending library of software and for downloading from BIONET. MacDNA is a small (65Kbyte), self-contained application that is fast and efficient at many straightforward DNA analysis tasks such as translations, character stripping, restriction mapping, consensus searching, inverting, formatting, etc. The program is not Mac-like, i.e., it uses no mouse or dialog boxes, but the current version will run under Multifinder. The current version has enhanced formatting and editing capabilities, and does single letter translations, not included in the original version.

OligoMutantMaker

OligoMutantMaker from Kevin Beadles of the Univ. of Calif. at San Francico has been made available to BIONET users. OligoMutantMaker simplifies the designing and screening of oligonucleotide-directed single amino acid substitution experiments by searching for nucleotide sequences which introduce a restriction endonuclease recognition sequence into the codon substitution site of the mutant. The program utilizes the redundancy of the genetic code to generate all possible nucleotide sequences for a given amino acid substitution (including nucleotide sequences in which silent mutations are introduced into the 5' and/or 3' codons immediately adjacent to the substitution site) and determines whether any restriction endonuclease recognition sequences are present. Any nucleotide sequence containing a restriction site is displayed or printed along with relevant information about the site such as its restriction enzyme(s), the random frequency of the enzyme's recognition sequence(s), the prototype of the enzyme, isoschizomers of the enzyme, and the unit cost of the enzyme from various biochemical suppliers.

OligoMutantMaker runs on the IBM PC and a version for the Macintosh is in preparation.

New Version of SEQAID II

SEQAID II is a multifunctional PC program for DNA and protein sequence analysis from Donald Roufa and Douglas Rhoads of Kansas State University. SEQAID II's functions include editing, extracting sequences from a GenBank floppy disk release, dot matrix comparison, fragment sizer, base composition, translations, protein structure and hydropathicity, restriction site search, and locating potential exons by codon bias.

One of the new features in release 3.0 is a gateway to the GenBank database (floppy disk release) and a user-friendly interface to search and extract genetic sequence data from the database. Included is a utility, GBHEADER.EXE, which permits users to modify the GenBank format file, HDR550.GBK, to accomodate future floppy disk releases. As implemented, SEQAIDII accesses floppy disk releases of GenBank either segmented on diskettes or combined in a subdirectory of the user's harddisk.

SEQAIDII version 3.0 also contains a new module that scans anonymous nucleic acid sequences for potential protein coding regions based upon codon usage frequencies. The module permits the user to construct and amplify required codon bias tables from appropriate nucleic acid sequences (cDNAs or spliced exon sequences) and to save the tables to disk.

IDEAS - As a sizeable minority of BIONET users had been requesting the availability of programs for protein structure analysis and prediction, last year BIONET investigated the availability of programs performing these types of analyses. It was decided that the IDEAS suite (Integrated Database and Sequence Analysis System) by Minoru Kanehisa of Kyoto University (and previously the National Cancer Institute) would be an appropriate package to obtain. An older version of the IDEAS package, ported to run on the BIONET DEC-20 computer, had already been made available by Dr. Kanehisa but it lacked the newer structural analysis programs needed. Since the newer version ran only under the VMS operating system on VAX computers, in February 1988 BIONET, in co-operation with IntelliGenetics, made available an IntelliGenetics VAX/VMS computer to BIONET users (see *Technological Research*). Accordingly, BIONET set up an account on an IntelliGenetics networked microVAX computer, and ported over the IDEAS package stripped of all programs save those pertaining to structure analysis or prediction and that would ensure an added functionality to BIONET. Complete on-line help files were written describing both the programs and the access procedures, so that BIONET users could easily review the functionality of these new programs before availing themselves of them.

The programs made available were:

- **STRALI**: The STRALI program performs a secondary structure prediction by homology alignment (Kanehisa, unpublished). It searches for local regions of similarity between a query protein sequence and one whose structure has been determined and deposited in the Brookhaven Protein Databank. The Kabsch-Sander secondary structure classification (see DSSP, below) is used to predict regions of secondary structure in the query.
- **CHOFAS**: CHOFAS is the popular Chou-Fasman (Chou,P.Y. and Fasman,G.D., Biochem.(1978) 47:251-276) algorithm for secondary structure prediction complemented by the Rose algorithm (Rose,G.D., *Nature*(1978)272: 586-590) for beta-turn prediction.
- **DELPHI**: DELPHI performs protein secondary structure prediction by Robson's method (Garnier et al.,J. Mol. Biol.(1978)120:97-120). The program makes predictions for helix, (extended) beta-sheet, reverse-turn, and coil. Decision constants, provided by the user, can affect the predictions greatly.
- **ALOM**: This program attempts to identify membrane proteins based on a discriminant analysis of hydrophobic amino acids (Klein,P. et al., in prep.). The program reports whether the sequence is likely to be integral or peripheral, together with the estimated likelihood.
- **HCOMP**: Displays on a graphics terminal the hyrophobicity profiles of two aligned protein sequences. The profile is calculated by smoothing the Nozaki-Tanford hydrophobicity scale. (Nozaki,Y. and Tanford, C. J. Biol. Chem. (1971) 246: 2211-2217).
- **HPLOT**: Plots the distribution of hydrophobic and charged amino acids using any of four possible algorithms: Nozaki-Tanford (Nozaki,Y. and Tanford, C. J. Biol. Chem. (1971) 246: 2211-2217); Hopp-Woods (Hopp,T.P. and Woods,K.R. Proc. Natl. Acad. Sci. USA (1981) 70: 3024-3828); Eisenberg (Eisenberg et al., Proc. Natl. Acad. Sci. USA (1984) 81: 140-144); or Kyte-Doolittle (Kyte,J. and Doolittle, R.F. J. Mol. Biol. (1982) 157: 105-132).

Since its introduction the IDEAS suite has been accessed an average of 23 times per month.

3.1.2.4 Data Contributors

BIONET has always provided rapid updates to all the major collections of sequence data including GenBank and EMBL nucleotide sequence collections, and the NBRF/PIR Protein Data Bank. This last year we have continued to expand the databases on-line that are related to molecular biology. This has often involved establishing contacts with database managers and providing them with the facilities on BIONET to maintain their data collections. The following summarizes our current and projected activities in this area:

Restriction Enzyme Database. We continue to provide the community with the latest additions to the Restriction Enzyme Database, through the cooperation of BIONET and Dr. Roberts at Cold

Spring Harbor. Modifications are mailed electronically to BIONET after they are incorporated into the on-line database at CSH. In addition, we provide the community with subsets of the list of enzymes that are commercially available. These subsets have been revised using the data in the July, 1988 version of the database. The lists are made available in IntelliGenetics format for use in the SEQ and PEP programs.

The catalog sources for each of the files are listed below:

amersham.lst	Amersham (8/87)
brl.lst	Bethesda Research Laboratories (6/88)
anglian.lst	Anglian Biotechnology Ltd. (1/88)
ibi.lst	International Biotechnologies Inc (7/87)
boehringer.lst	Boehringer-Mannheim (5/88)
neb.lst	New England Biolabs (7/88)
pharmacia.lst	Pharmacia P-L Biochemicals (5/88)
promega.lst	Promega Biotec (9/87)
usb.lst	United States Biochemical Corporation (4/88)

These files are extremely useful since the full Roberts' database now consists of over 900 enzymes and most users are only interested in enzymes which they can readily access. Users can also create custom databases of enzymes using the information in the commercial enzyme lists to limit their analyses to enzymes in their lab or from their favorite suppliers. Analyses run with these shorter enzyme lists will, of course, also be correspondingly faster.

SWISS-PROT. Since June, 1987 the BIONET staff has made the SWISS-PROT protein sequence database from Amos Bairoch available on BIONET. SWISS-PROT contains data obtained from the NBRF/PIR database, data translated from the EMBL DNA sequence Data Library, as well as sequences entered in-house. To maintain SWISS-PROT on BIONET requires conversion to both the IntelliGenetics format and to the FASTA format for searching by the FASTA-MAIL program. In addition, testing is done after the conversions are made to insure the integrity of the data. Maintaining the original, IntelliGenetics, and FASTA formatted versions of SWISS-PROT currently requires over 28 Mbytes of disk storage. When first released on BIONET, SWISS-PROT release 4.0 contained 1,036,010 residues in 4387 sequences. Since then the databank has approximately doubled; the current release 8.0 contains 2,224,465 residues in 7724 sequences.

Genetic Variations of Drosophila melanogaster - the "Red Book". BIONET has long had the full text of Lindsley and Grell's classic work "Genetic Variations of Drosophila melanogaster" available on line. This was made possible by help from the author, Dan Lindsley and with the permission of the original publisher, Carnegie Institution of Washington. Dan Lindsley holds an NLM grant to update this book and has also been forwarding chapters of the book to BIONET as they are finished. Because of this, the most complete collection of Drosophila mutants are always available on the BIONET computer. This book can be searched using one of several different text searching programs (QUEST, FIND and XSEARCH) all of which allow searches for complex boolean relationships between search terms. This also permits BIONET users to search for all genes by name, by phenotype, by affected tissues, by genetic location and by cytological location. This size of the text is still small enough to permit a complete serial search in a matter of seconds. This year BIONET received and put on-line the revisions of the chapters describing loci and mutant alleles whose names begin with the letters A through R.

SV40 Mutant List. On July 15, BIONET made available a contributed database of the SV40 large T antigen mutants compiled and maintained by Dr. J. M. Pipas of the University of Pittsburgh. The database consists of five lists of mutants divided into the following groups: early region deletion mutants including either sequence or map information; early region point mutants including either base-pair change (and corresponding amino acid substitution) or map position; and the primary sequence of truncated T antigen mutants. Each list is annotated and has a complete set of literature references which describe the mutations listed in detail. Since its introduction on BIONET the database has been distributed to five geographically distinct regions in the U.S. and in Europe using BIONET's existing ARPANET and BITNET connections, as well as being used at the rate of 16 references per month on BIONET itself.

LiMB Database. In February, BIONET made available on-line the LiMB (Listing of Molecular Biology databases) database, created by the staff at Los Alamos National Laboratories. LiMB contains information about the contents of databases related to molecular biology as well as details of how they are maintained. It was created to facilitate the process of locating and accessing databases that the research community depends on; it is also of use to those who are doing research in designing and linking these databases. Information for each database includes the purpose of the database, contact addresses including network addresses for the database staff, the history of the database, source of the database data, literature references, cross-references to other databases, details of the computer-readable form of the database, and the size of the database.

Removal of Brookhaven database from BIONET. This year access to the Brookhaven structural database had to be curtailed due the the imposition of an $8,500 per year fee for network access by Brookhaven. Apparently a fee of this magnitude was necessary because the database effort is largely funded through the sale of tapes. The database was continued only for in-house research by the BIONET staff at the old fee of only about $200 per release. The low usage of the database on BIONET due to the lack of graphics capabilities on the system precluded the expenditure of $8,500. Only the DSSP program which extracted secondary structure information was available for use on the system. Given different circumstances in the future the database may be restored to BIONET. However the expenditure of such a large sum of money for this database compared to the costs of other molecular biology databases (about 10x more expensive) could be avoided if more funds were provided by the government to Brookhaven to finance its operation.

Protein Crystallography Directory. Dr. M. M. Teeter of Boston College maintains a list of the electronic mail addresses of all protein crystallographers in the U.S., Canada, Europe, and elsewhere. A copy of the latest version of the list is periodically sent electronically to BIONET over BITnet, where it is made available to any BIONET users who wish to browse it using a free-form text searching utility called FIND available on BIONET. All a user need do is attach the appropriate network suffix (e.g., .bitnet, .janet, or .earn) in order to send electronic mail to any of the crystallographers listed in the database.

3.1.2.5 Liaison with Other Resources

Several accounts have been established on BIONET to promote interaction with other, related Resources. The following is a summary of current sites with which we can exchange information.

BRTP Mailing List. Following the February 1988 meeting at the DRR BIONET established an electronic mailing list for the Biomedical Research Technology Program. By sending a single mail message to the address brtp@BIONET-20.bio.net it is now possible for scientists at any of the BRTP-funded resources to communicate with scientists at all other resources.

Molecular Biology Computer Research Resource. The MBCRR, at the Dana-Farber Cancer Institute at Harvard, shares information through mail delivery via the GENE account and via the bulletin board system. An MBCRR bulletin board is available on the DEC 2065. Dr. Jurka at BIONET has also continued his work on the MASE editor in collaboration with Mr. Donald Faulkner at the MBCRR.

Molecular Biology Information Resource. The MBIR at Baylor formerly communicated with BIONET through Dr. Lawrence's account on the BIONET computer. After having established an Internet connection last year, the MBIR has been on our list of BIONET newsgroup recipients.

Protein Identification Resource. BIONET has provided a bulletin board expressly for the PIR which now allows members of that database staff to communicate easily with users around the world. In addition the PIR along with GenBank and EMBL now receives automatic electronic data submissions via BIONET's XGENPUB program using the common data submission form.

GenBank. GenBank continues to utilize the GenBank bulletin board on BIONET which now has worldwide distribution through the BIOSCI bulletin board network. GenBank also receives new sequence data submissions from BIONET's XGENPUB program. XGENPUB was released in 1987 and has been used to submit a total of 202 new sequences to the databases; 166 during this reporting period (12/87 - 11/88) for an average rate of 14 per month.

EMBL Databank. The EMBL continues to utilize its bulletin board on BIONET to communicate with the scientific community. It also receives electronic data submissions from BIONET's XGENPUB program as noted above.

Pittsburgh Supercomputer Center. From August 8 - August 12, Spencer Yeh, the BIONET Applications Analyst, participated in the Biomedical Supercomputer Workshop at the Pittsburgh Supercomputing Center (PSC). The Pittsburgh Supercomputer Center utilizes a Cray Y-MP supercomputer and will be the first supercomputer center to receive a Cray 3 computer when they become available in 1990. In addition, the PSC has a three-year $2.2 million grant from the NIH's Division of Research Resources to provide the biomedical community with supercomputing resources, training, and user support. Because of this, the PSC already has both the GenBank and PIR databases on-line. The purpose of the visit to the PSC was to establish the feasibility of providing BIONET users remote access to the PSC's Cray computer for computationally-intensive tasks.

The workshop covered topics such as using the VAX frontends to the Cray, using existing software at the PSC for biomedical research, and optimization/vectorization techniques for vectorizing FORTRAN code on the Crays. Discussions with Hugh Nicholas and David Heerfield, Scientific Specialists at the PSC, addressed important issues such as the network connectivity of the PSC for remote job submission, expected speed improvement on typical BIONET applications when ported to the Cray, time allocation schemes for Cray CPU usuage, suitability of the Cray for languages other than FORTRAN, especially C code, and the projected programmer time required for porting existing applications to the Cray environment. It was learned that the PSC is changing to a UNICOS (Cray UNIX) operating system which will facilitate systems development efforts since the BIONET systems programmers are already conversant with UNIX systems.

As a result of the discussions it was concluded that it would be feasible to allow BIONET users remote access to the PSC's Cray. However, important issues still need to be addressed such as the projected speed and reliability of the NSFnet over the next 5 years, network and Cray queuing delays for remote jobs, and the overhead involved in submitting jobs to a supercomputer. While the Cray can achieve speed increases of 10 - 100 times the performance of a SUN 3/280, the turnaround time will probably be limited by the network response from California to Pittsburgh, and by the queuing system at the PSC for batch jobs. Jobs which require less than 30 min. - 1 hr. of Sun CPU time are probably too small to warrent the overhead of remote job submission to the PSC. The workshop provided an excellent introduction to using Cray computers, and the information gained will allow BIONET to evaluate its options for handling computationally intensive tasks.

Biological Matrix Workshop. BIONET has continued to support the Matrix project. Dr. Jurka, the BIONET scientist, attended the last Matrix meeting in Washington in October and BIONET has continued to maintain a newsgroup for the Matrix project.

BIOSCI bulletin board network distribution centers. BIONET initiated collaborations with five other university sites to establish the BIOSCI bulletin board network. This is described in more detail below.

Journal Editors on BIONET. Last year BIONET established accounts for the editorial board of the *Journal of Biological Chemistry*, *CABIOS*, *Cell*, and the Washington office of *Nature*. This year an account has been opened for the *Journal of Bacteriology*. These accounts serve several purposes. First, they allow easy communication between the scientific community and the on-line editorial staff. Second, they allow the editors access to sequence analysis software which they may use in reviewing manuscripts. Third, the accounts increase awareness among editors of the advantages of electronic networking, especially in regards to the problem of data submission to the nucleic acid and protein sequence databanks. Although the usual reaction on the part of journal editors has been one of reluctance to become actively involved in the data submission process (understandably so in light of their heavy workloads), their involvement in BIONET alerts them to the availability of on-line data submission software (the BIONET XGENPUB program described above) and they can then pass this information on to other scientists.

This year Dr. Herb Tabor was trained in the use of the system by Dr. Kristofferson while in Washington for the February DRR meeting. Dr. Kristofferson also trained one of Dr. Tabor's assistants at the *J. Biol. Chem.* offices in the use of electronic mail and file transfers. Joseph Palca of the Washington office of Nature has actively used the system during the course of the year for electronic communications. Recently the *Journal of Bacteriology* has been provided with an account on BIONET for the purpose of advance publication of the table of contents of that journal. BIONET will establish an electronic bulletin board for this purpose.

Training of the editorial staff in the use of BIONET can usually be accomplished by direct terminal links with simultaneous verbal instruction over the telephone. In the course of about 45 minutes a person can become proficient in using the electronic mail and bulletin board facility and also learn the use of the on-line help system. This allows them to explore other features of the system at their leisure. The BIONET consultants are also available to assist the editors when called upon.

3.1.2.6 Bulletin Boards and Leaders

The following bulletin board topics are currently available on the system.

Bulletin Board Name	Description
AGEING	Scientific interest group
ASK-BIONET	User queries and consultant responses
BIO-CONVERSION	Scientific interest group
BIO-MATRIX	Applications of computers to biological databases
BIONEWS (BIONET-NEWS)	General announcements
BIOTECH	Biotechnology issues
CONTRIBUTED-SOFTWARE	Information on programs contributed by users
EMBL-DATABANK	Communications about EMBL databank concerns
EMPLOYMENT	Job openings
GENBANK-BB	Communications about GenBank database matters
GENE-EXPRESSION	Scientific interest group
GENOMIC-ORGANIZATION	Scientific interest group
INFO-1100	Computer interest group
INFO-AILIST	Computer interest group
INFO-AMIGA	Computer interest group
INFO-ATARI16	Computer interest group
INFO-IBM-PC	Computer interest group
INFO-KCC	Computer interest group
INFO-KERMIT	Computer interest group
INFO-LAW	Assorted legal information
INFO-MAC	Computer interest group
INFO-MODEMS	Information abooout modems
INFO-NEURON	Neural network computing
INFO-SUN-SPOTS	Computer interest group
INFO-TELECOM	Telecommunications
INFO-VAX	Computer interest group
MBCRR	BBoard for MBCRR announcements
METHODS-AND-REAGENTS	For reagent exchanges and announcements about lab methods
MOLECULAR-EVOLUTION	Scientific interest group
ONCOGENES	Scientific interest group
PC-COMMUNICATIONS	Information on communications software
PC-SOFTWARE	General PC software announcements
PIR	Messages to and from the PIR database staff
PLANT-MOLECULAR-BIOLOGY	Scientific interest group
PROTEIN-ANALYSIS	Scientific interest group
RESEARCH-NEWS	General interest items about science
SCIENCE-RESOURCES	Information about funding agency policy, etc.
SWISS-PROT	Messages to and from the SWISS-PROT database staff
YEAST-GENETICS	Scientific interest group

The leaders of the indivdual boards are:

AGEING	Sydney Shall
ASK-BIONET	David Kristofferson
BIO-CONVERSION	Eng-Leong Foo
BIO-MATRIX	Dan Davison
BIONEWS	David Kristofferson

BIOTECH	Deba Patnaik
CONTRIBUTED-SOFTWARE	BIONET Staff
EMBL-DATABANK	Graham Cameron & David Kristofferson
EMPLOYMENT	David Kristofferson
GENBANK-BB	Christian Burks & David Benton
GENE-EXPRESSION	Bill Sofer
GENOMIC-ORGANIZATION	Tom Marr
MBCRR	Susan Russo
METHODS-AND-REAGENTS	David Kristofferson
MOLECULAR-EVOLUTION	Dan Davison
ONCOGENES	David Steffen
PC-COMMUNICATIONS	David Kristofferson
PC-SOFTWARE	Doug Brutlag
PIR	David George
PLANT-MOLECULAR-BIOLOGY	Robert Jones
PROTEIN-ANALYSIS	Amos Bairoch
RESEARCH-NEWS	Sunil Maulik
SCIENCE-RESOURCES	Michele Cimbala
SWISS-PROT	Amos Bairoch
YEAST-GENETICS	John Thompson

Note that the INFO- bulletin board material is received from information sources outside of BIONET.

This year was a particularly exciting time in the development of the BIONET bulletin board system as it became the nucleus of the new international BIOSCI bulletin board system.

Because scientists work on a variety of computer networks around the world, BIONET recognized the necessity of developing a mechanism to allow all of them to communicate without the necessity of learning the peculiarities of accessing each network. We sought out computer sites on all major international networks and arranged to have parallel copies of the original BIONET bulletin boards accessible from these sites. Besides BIONET in the U.S., other major BIOSCI distribution sites are now situated at the SERC laboratory in Daresbury, England; the University College, Dubulin, Ireland; the University of Uppsala in Sweden, and the University of Helsinki in Finland. Recipients of the bulletin boards from these sites are located around the world from New Zealand and Australia, the Far East, and Israel, throughout Europe, and back to North America. The bulletin boards are available to users on the ARPANET (from BIONET), BITNET (from BIONET), EARN (from Dublin, Daresbury, Helsinki, and Uppsala), Usenet (from BIONET), NSFnet (from BIONET), and JANET (from Daresbury). Users in any particular location need only post or receive messages from their closest site. Any postings at any one site are automatically forwarded by the central BIOSCI sites to all other participants on all of the above-listed networks.

A copy of the BIOSCI information sheet mailed electronically to people who request information is provided in *Appendix V.*

The following list contains the number of messages posted to each BIOSCI board from 12/87 through 11/88. All told 956 messages were posted which was an increase of 74% over the previous year. As many molecular biologists are only now discovering the use of electronic mail and bulletin boards we expect that these high growth rates will continue into the foreseeable future. For the last five years BIONET has looked on effort expended on electronic communications as a long-term investment. This year it is clearly starting to pay off.

Bulletin Board	Messages Posted
AGEING	0 (being established)
BIO-CONVERSION	6 (started end of 11/88)
BIO-MATRIX	46
BIONEWS	187
BIOTECH	118
CONTRIBUTED-SOFTWARE	21
EMBL-DATABANK	17
EMPLOYMENT	93
GENBANK-BB	33
GENE-EXPRESSION	14
GENOMIC-ORGANIZATION	0
METHODS-AND-REAGENTS	105
MOLECULAR-EVOLUTION	44
ONCOGENES	12
PC-COMMUNICATIONS	12
PC-SOFTWARE	57
PIR	18
PLANT-MOLECULAR-BIOLOGY	5
PROTEIN-ANALYSIS	33
RESEARCH-NEWS	72
SCIENCE-RESOURCES	51
SWISS-PROT	4
YEAST-GENETICS	8

3.1.3 Technological Research

3.1.3.1 Research Efforts by the BIONET Scientist

The activities during 1988 can be divided in the following three groups: (1) Studies on interspersed repetitive elements with emphasis on Alu and L1 subfamilies; (2) Design and development of software for sequence analysis (with emphasis on sequence extraction, multiple alignment and classification); (3) Collaborative research on evolution of Alu and protein sequences (discussed above under **Collaborative Research**).

1. The biological findings on Alu classification have been published (*Proc. Natl. Acad. Sci. USA 85*, 4775-4778, 1988 & *Nucleic Acids Res. 16*, 766, 1988 see *Appendix I*). In addition to the work on Alu sequences, classes of KpnI sequences have been discovered. A manuscript on this subject will be submitted by December 1988.

2. The following activities occurred: (a) Development of a sequence editor in collaboration with Donald Faulkner of Harvard (*Trends in Biochemical Sciences 13*, 321-322, 1988). After publication of the article, several new functions have been added to MASE upon our suggestions: enhancements on COLUMN-CORRELATION, CREATE-LOCUS, SIMILARITY-DISCARD-GAPS, RENAME-LOCUS, JUMP-TO-POSITION-ABSOLUTE, modifications of PATTERN-HIGHLIGHT, MODE-ALIGNMENT, MODE-DNA, MODE-PROTEIN. In addition, we assisted in testing of these and other functions added to MASE during last year; (b) In-house work on a sensitive algorithm for sequence classification with emphasis on Alu and KpnI repeats. Previous attempts by other investigators to classify Alu repeats based on overall sequence similarities proved unsuccessful (Bains, 1986). This was quite inevitable since the diagnostic

positions are only a small portion (1-6%) of the total number of bases in Alu sequences and the distinction between them and the statistical noise could not easily be made using standard tree analysis.

3.1.3.2 FASTA-MAIL

Because of the increasing demands on the DEC 2065 BIONET needed to find alternative computing resources. Towards the end of last year, Sun Microsystems generously donated a new central computing facility to BIONET. Although some software development is necessary before we can transfer users directly to the Sun system we were able to use it remotely fromj the DEC for database searches through the use of our new FASTA-MAIL program. Much of the Dec's CPU resources were being spent on genetic sequence library searches, which could take up to six hours of CPU time. In an effort to ease CPU usage, we undertook to write an interface to FASTA, a sequence search program written by William Pearson that runs under UNIX. The interface we devised is known as FASTA-MAIL.

FASTA-MAIL will take submissions received via Internet mail, queue them for batch processing, and mail the results back to the submitting address. On the DEC 2065 users run a simple program and answer a few questions about the type of search that they wish to perform. This program submits these instructions together with the query sequence data to the Sun 3/280 computer at BIONET. Currently, there are two queues on our Sun that run concurrently to handle protein and nucleic acid searches, respectively. The batch processing program processes a limited grammar at the beginning of each message so that it may pass command line arguments to FASTA.

On the Sun end a message is received by the mail processing agent known as **sendmail**, which delivers the message to a program instead of a user. In this case, the program is one that checks whether the user has requested a nucleic acid search or a protein search, and places the message in the appropriate batch queue. It should be noted that as initially implemented, the batch queues used were, in fact, printer queues because SunOS's batch facilities were too weak to support multiple queues. Sequence data would enter through the print spooling system and be processed as printing output until the output filter was called. At that point, we specified our own output filter, a program known as *fastaq*. *fastaq* would then compare the sender of the database query with a list of authorized users. Once validated, the message headers would be stripped (although the program retained the return address), and a small set of parameters are read in - ktuple and database name. The rest of the data is then passed to FASTA, whose output is sent directly to the UNIX mail program.

Besides allowing access to BIONET users on the DEC 2065 computer, the FASTA-MAIL program ushers in a new means of performing database searches. The program can be easily modified to accept input from **any electronic mail site in the world**. Soon we expect to allow users at other sites access, but outside jobs will be placed in a lower priority queue so as not to impact significantly on our registered users. Expansion of the computer resources available at BIONET would allow us to serve a far greater number of users **very economically**.

3.1.3.3 Development work on the new Sun central computing facility

In June of 1988, the six Sun 3/60M computers and one 3/280S computer donated by Sun Microsystems arrived at BIONET. These computers are nearly operational and are providing some service to the BIONET community, particularly via FASTA-MAIL as noted above. In addition to many smaller projects (not listed), the following work on the new system has been accomplished:

- The BIONET/BIOSCI mailing lists that were administered on the DEC-20 are now being administered on the 3/280S. This will take some of the load off of the DEC-20's already loaded mailer.
- The BIONET Sun has become the major distribution point for the BIONET/BIOSCI mailing lists, passing them onto other major Internet sites.
- As mentioned previously, FASTA-MAIL was written and implemented in August so that genetic sequence searches could be done on the Sun.
- A hierarchal help system has been implemented, and the BIONET documentation is being moved into this system.
- *newuser*, a program developed by one of our staff is being implemented so that users will configure their environments the first time they log into the Suns.
- *MM*, a mail management system has been installed. An almost identical mailer was run on the DEC-20. This will eliminate relearning the electronic mail program when users migrate to the Sun system.
- *netnews*, an electronic bulletin board system, has been installed.

We shall be begin moving users to the Suns from the DEC-20 as soon as accounting software is in place.

3.1.3.4 BIONET Satellite Software for VAX/VMS systems

The BIONET Satellite software package provides the mechanisms for VAX/VMS computer systems to exchange electronic messages with a variety of computer installations. Messages are sent using telephone connections in a standard electronic mail format. Messages delivered are compliant with the address conventions of RFC822 (the standard for the format of Internet text messages). Messages can be directed to individuals or sent to a bulletin board facility. The bulletin board system implements the Standard for Interchange of USENET messages. This standard allows the host system access into the USENET news network. The software components consist of two integrated subsystems which implement the functionality as described above. The Pascal Memo Distribution Facility (PMDF) implements the MMDF protocol which provides the interface for message transmission. NEWS, written by Geoff Huston is the system which provides general conferencing in the form of a bulletin board service on the VAX/VMS host system.

The BIONET satellite system was originally designed to connect to a DEC 2060 computer running the TOPS-20 operating system. Message transmission to VAX/VMS systems was implemented using a non-standard communcations protocol, CAFARD. This protocol would then interface with a mail delivery system called Pony Express. Pony Express software is a proprietary product of SRI. The acquisition of a SUN 3/280 as the primary host system for BIONET has necessitated changes in system software. The SUN machine uses the UNIX operating system. To support the CAFARD and Pony Express systems on the SUN machine would have required extensive development and maintenance time. The alternative was to look for public domain software which would provide the original functionality of the Satellite software. The selection of PMDF and NEWS were selected, in part, to realize these goals. The MMDF protocol has been used extensively on UNIX machines and the PMDF system provides compatibilty on all current VAX/VMS systems. The USENET network has been used for years as a way to conference information between end users. The NEWS system provides both connectivity and a user interface to USENET. These two subsystems are in the public domain and have souce code that is distributed. This new version of the Satellite software is now undergoing local testing. Installation and documentation procedures for VAX/VMS sites are near completion. It is anticipated that site testing will begin in early January 1989.

3.1.3.5 The RICH program

It is now known that the amino acid composition of a protein plays a major role in determining its folded state (Sheridan RP *et al.* (1985) *Biopolymers 24*: 1995-2003). A fundamental pattern-matching problem in protein sequence analysis is that of finding the largest subsequences given certain density (composition) criteria only. For instance, one may wish to find the largest region in a 500 amino acid protein that contains >20% proline and >30% cysteine residues. A related problem may be finding all protein subsequences in a database that have greater than 60% hydrophobic residues.

As a result of requests such as these from BIONET users, an algorithm has been developed that will scan a sequence and heuristically discover the largest subsequence(s) that satisfy any given density criteria. Current algorithms for finding rich regions in sequences (such as RICH in the SEQ program; Brutlag, D., Clayton J., Friedland, P., & Kedes, L.H. (1982) *Nucl. Acids Res. 10*: 279), find the first such subregion satisfying the density criteria and then expand in increments of one unit (base or amino acid) until the combined density falls below the density threshold. Following this, the region is reported, and the algorithm proceeds to find the next such region. The fundamental problem with algorithms of this type is that they fail to look far enough ahead (or back) to see if sufficiently dense regions could be incorporated with the current one in order to locate the largest subsequence. Our algorithm uses a heuristic search procedure to find all seed regions satisfying the density criteria and then attempts to link all such "seeds" together. By finding all seeds in the initial pass, the algorithm circumvents the problems of the other implementations and produces the largest subregion in the second (or following) iteration or optimizing pass.

The implementation of this algorithm, termed RICH, is near completion and will be available on BIONET in 1989. The implementation has been optimized to allow RICH to scan entire databases such as the NBRF-PIR protein sequence database and locate patterns describable by their density characteristics. Uses of the RICH program might include finding hinge structures in immunoglobulin sequences (known to be rich in prolines and cysteines; Huber, R. and Bennett, W.S. (1987) *Nature 326*: 334-335) or verifying that a protein sequence satisfies the PEST hypothesis (Rogers, S., Wells, R., & Rechsteiner, M. (1986) *Science 234*: 364-368), i.e., if its half-life is related to the density of P (proline), E (glutamic acid), S (serine), and T (threonine) residues. A preliminary search of the NBRF-PIR database using RICH has shown surprising results in the type of sequences containing one or more dense PEST regions, and further studies to verify and quantify this effect are progressing. A manuscript describing RICH is in preparation, including examples and results arising from using RICH to scan databases for density-dependent patterns.

3.1.3.6 BioCard - a prototype menu-driven interface for BIONET

BIONET had previously investigated using application toolkits to develop graphics-based, menu-driven interfaces to terminal emulators running locally on users' personal computers. The aim of these investigations was to create an interface containing all the information about BIONET as easily accessible, cross-referenced help text, and to allow users to activate software or database search queries on BIONET using simple switches/buttons which would then run command-files on the central BIONET computer. With the advent of HyperCard for the Apple Macintosh, an information management system with an built-in applications generator, the creation on icons, menus, "dialog" boxes, windows, and hypertext tailored specifically to BIONET users has become straightforward. The first HyperCard application for BIONET, termed BioCard, consists of a set of graphics screens containing icons for all of the information currently on BIONET's HELP ME system, as well as buttons representing the most commonly performed tasks on BIONET. A user of BioCard could have it automatically dial the nearest Telenet or CompuServe phone number, connect to BIONET, and then run a sequence analysis program on the BIONET computer. The user can also switch between remote (BIONET) mode, and local (BioCard) mode, thus allowing users to browse through HELP information without having to quit their current application.

Currrently, BioCard consists of the complete HELP ME system of BIONET in an easily searchable Hypertext format, as well as twelve buttons which, when activated, run the most commonly automated sequence analysis functions on BIONET such as database similarity searches, restriction mapping, sequencing gel assembly, etc. BioCard is also bundles with MacKermit, a public domain terminal emulation package for the Macintosh from Columbia University. Under the Macintosh's MultiFinder operating system, users can switch between local (BioCard) mode and remote (BIONET) mode in a simple and intuitive manner.

One advantage of Hypercard over other applications generators is that it is easily customizable by the user, even one with only a rudimentary familiarity with computers. Thus BIONET users will have the opportunity to select those icons, windows, and buttons that they require and use most frequently, while discarding others of no interest to them. Further, new functionality may be created by coupling or modifying the existing application icons. Finally, adventorous users may wish to design their own interfaces using BioCard as a model. Thus BioCard serves as an opportunity for BIONET to explore the many human factors engineering aspects of interface construction. Once these factors are known (through feedback from the user community over the BIONET network), BIONET plans to develop hardware-independent user interfaces using such windowing protocols as X-Windows from MIT, a public domain windowing system that is finding acceptance with such personal computer/workstation vendors as IBM, DEC, Sun, and Apple. A preliminary version of BioCard is under testing, and a first release should be available in 1989.

3.1.3.7 XGENPUB

One of the original goals of BIONET was to aid in several of the database efforts including GenBank, EMBL and the Protein Identification Resource. Initially we felt that a major contribution that BIONET could make would be to make these databases more readily available by providing software tools for database searching and analysis. However it became clear that a great number of DNA sequences were actually being determined using the IntelliGenetics GEL program on the BIONET computer. We felt that BIONET could provide a further service to both the community and to the database efforts by developing software that would allow the scientist to annotate his sequence according to the standard GenBank format and mail the sequence and its annotation to GenBank electronically. In 1987, BIONET completed work on and released the initial version of GENPUB, currently named XGENPUB on the DEC-2065. The GENPUB program is a forms-oriented display editor that allows a person to fill in a template based on the GenBank submission form (and which can be readily changed if the GenBank form changes) giving all the requisite data about a sequence. The program automatically inserts the sequence in the appropriate place in the form by copying the sequence from a designated file on the BIONET computer. When the form is completed a single keystroke forwards the information to both the GenBank computer at Los Alamos and the EMBL computer in Heidelberg for inclusion in the next issue of the databases. At that point the entry is verified by the GenBank and EMBL staff and if they have questions about the data they can query the author by electronic mail at BIONET. To date, GENPUB has been used on BIONET to submit 202 sequences to the databases.

This year some new features were added to the XGENPUB program:

- the ability to accept any sequence data file regardless of format;
- the ability to preview the data submission form inside the XGENPUB program prior to going into the editor and;
- additional internal help documentation.

The PIR database was also added along with GenBank and EMBL as a recipient of data submissions after a common data submission form was adopted by all three databases.

Direct electronic submission mediated by GENPUB eliminates many errors in transcription. GENPUB takes sequence files directly from the software used to determine the sequence and submits it to the database. If local PC software is used to determine a sequence, the data can be sent to BIONET using an error-checking protocol such as Kermit or Modem and forwarded to GenBank via GENPUB eliminating all transcription errors. More importantly, GENPUB recruits the scientist determining a sequence to annotate it, eliminating the problems of reading and interpreting the publication and further simplifying the job of sequence collection. This concept of recruiting the aid of the scientists performing sequencing to help build the database was fundamental to the IntelliGenetics application for the GenBank contract. Since that contract was awarded to IntelliGenetics, GenBank has undertaken to make an even more sophisticated form of this program which will include error checking on the data entered into the annotations and which will be extremely portable, running on a number of microcomputers as well as mainframes. When this program is distributed to scientists who are developing new sequences, it will markedly increase the rate and quality of the data entering the GenBank database. BIONET is proud to have served the community in this way and looks forward to even closer working relationships between itself and the database efforts. We have taken an important role in not only making databases more accessible, but have also taken an active role in helping accumulate the data as well.

3.1.4 BIONET Training Program

This year we have continued holding intensive training sesions in our in-house training facility and have also given a number of demonstrations and lectures at outside sites.

We conducted two day, in-house training sessions in March, May, July, September, and November.

These five courses served a total of 45 BIONET users. The courses continue to stress the integration of the programs to solve problems. These problems include sequence entry and editing, sequencing gel management, nucleic acid and peptide sequence analysis, database structure and sequence retrieval, collecting sequences, pattern searching, and sequence similarity searches. In November we expanded the training program to include an evening session on connecting to the BIONET computer, transferring files, using a text editor, addressing electronic mail, and accessing other BIONET programs. This additional session will be included in future in-house trainings, as well. The schedules for each training are included in *Appendix VI*. The training material for these sessions has also continued to be revised by the staff during the course of the year. A new training manual for use in the courses has been produced and continually updated.

Outside seminars and demonstrations were held at Rutgers University, the NIH, the Protein Society meeting, the FASEB meeting, the Annual Meeting of the Canadian Society of Microbiology in Windsor, Ontario, and at Ohio State University at Wooster. A lecture is also planned next February in San Francisco at the annual ASCB meeting. In addition to these events, Dr. Kristofferson helped train the some of the journal staff at the offices of the *Journal of Biological Chemistry* and the Washington office of *Nature* as described above under **Liasons with Other Resources**.

Rutgers. BIONET presented a poster at the bi-annual meeting at Rutgers University's workshop titled "Computers in Molecular Biology" at the Waksman Institute/Center for Advanced Biotechnology and Medicine, New Brunswick, New Jersey, April 13-15. Approximately 50 people attended. The poster described new developments at BIONET and included hands-on demonstrations of using BIONET to solve molecular biology computing tasks. In addition, BIONET participated in a panel discussion on the future of molecular biology computing. Fourteen BIONET applications were obtained as a result of the meeting.

NIH Users' Group Meeting. BIONET, in association with IntelliGenetics, organized a User's Group meeting at the National Institutes of Health, Bethesda, Maryland on August 25. The meeting featured talks describing the BIONET resource and some of the software available on it. In depth presentations were also given on sequence similarity searches and alignments, assembling and managing sequencing fragments, and recent changes in the IntelliGenetics software package. A total of 70 people attended the all-day session.

Protein Society Meeting. BIONET was invited to present a poster and demonstration at the "Computer Workshop on Protein Analysis Software" at the Second Annual meeting of the Protein Society in San Diego, August 13-17, 1988. Over 1000 attendees were expected to be present for this meeting. Two posters were presented, titled " Protein Databases and Analysis Software on BIONET" and "Locating Amino Acid Patterns by Composition" that described the current state of research and technological development in protein analysis software at BIONET. In addition during the 3 hour workshop over 30 applications were obtained and 50+ people were given hands-on experience at logging into BIONET over the Telenet/CompuServe networks and making use of the BIONET software.

FASEB Meeting. The BIONET Applications Analyst, Spencer Yeh, attended the annual meeting of the Federation of American Societies for Experimental Biology (FASEB) in May. A portable computer was used to demonstrate the BIONET system to meeting attendees. Use of the new

Tektronix graphics display in the Kermit communications program was illustrated by modifying circular plasmids within the CLONER program. In addition to generating 39 requests for new BIONET applications, the many current BIONET subscribers attending the meeting were able to obtain personal assistance with analyzing their data on BIONET. Suggestions for improving the Resource and general comments about the usefulness of the Resource to the laboratory biologist were also conveyed to the BIONET staff.

Canadian Society for Microbiology. Dr. David Kristofferson was an invited speaker to address the members during the opening symposium of the Canadian Society for Microbiology annual meeting. The meeting occurred on June 20th in Windsor, Ontario, and the opening session, attended by over 100 scientists, was entitled "Computers in Microbiology." The features of the BIONET system were described and many stimulating questions and discussions followed the presentation.

Ohio State, Wooster. The Ohio State University at Wooster conducted a summer workshop on DNA sequencing and cloning techniques attended by 27 scientists from throughout the Mid-West. This workshop included sections on computer-aided sequence analysis. Dr. Kristofferson addressed the group on June 21st. He also gave an on-line demonstration of BIONET following the lecture during which several users acquired hands-on experience and had their questions answered about the use of the Resource.

3.1.5 Resource Facilities

Previous reports have discussed the DEC-2065 and the various software and database libraries provided by the BIONET Resource. In this section we highlight significant changes and additions to the suite of hardware and software that comprise BIONET.

3.1.5.1 BIONET/SUN Agreement

Because the DEC-2065 is rapidly becoming obsolete and because equipment funds have been limited, in the summer of 1987 we submitted a proposal to Sun Microsystems to obtain a new central computing resource for BIONET. Sun agreed to donate a central Sun 3/280 file server and six Sun 3/60 client workstations. This equipment arrived in spring and became operational in June. This BIONET initiative has saved the NIH $150,000 in equipment costs! Development efforts on the new system were described above under **Technological Research**.

3.1.5.2 Computer Hardware and Telecommunication Networks

Hardware.

The current BIONET Central Resource Machine is a Digital Equipment Corporation 2065 computer with an NI20 ethernet network interface. The ethernet interface provides access to the ARPANET and other IntelliGenetics' resources.

The hardware configuration is as follows:

```
KL10-E Model R Processor:

      2 MF20/MG20 Memory controllers
   3 MW MG20 Memory
        MCA25 Cache Buffer Memory
```

2 RH20 Massbus Channels
NI20 Ethernet Interface

Console and Front End Processor:

PDP-11/40 CPU, 32 KW 16 bit memory
RX02 Dual floppy disk drives
8 DH11 Terminal interfaces   8 * 16 TTY lines each = 128 lines
RH11 Massbus Channel
LP20 Line printer interface

Peripherals:

3 RP07 disk drives   111MW each
RP06 disk drive   39MW
372 MW Total disk storage
TU78 1600/6250-BPI tape drive
LP26 600 LPM Line printer

Disk space (data storage)

Public structure (PS:) disk space use on the 2065 is dynamic. The following snapshot is representative of typical usage, and is taken from December 1988.

Total disk space	433,000	(pages--222 million words)
Overhead/Common	<150,000>	(Core, System and System Support Libraries)
Swapping Space	< 25,000>	
File system Overhead	< 88,000>	(Directories and index pages)

	170,000	
BIONET Allocation	153,000	(90% of the available space)
BIONET Usage 12/88	<155,000>	

Unused space	< 2,000>	(Available for BIONET growth)

We conclude from these figures that BIONET is currently using 8,000 pages more than its disk space allocation.

Sun Computers

In May of 1988, BIONET took delivery of a substantial donation of computing hardware from Sun Microsystems. The BIONET staff, especially, Eliot Lear, one of the BIONET Systems Programmers, has been working to configure the hardware and software on the new Suns, so that we may gradually move users off of the Dec-20 and on to the new Sun computers.

The current Sun hardware configuration is as follows:

Sun Microsystems 3/280S Data Center Server

24 MB ECC Main Memory
1 Xylogics 753 3.0 MB/second SMD disk controller
1 Xylogics 472 tape controller
1 Systech MTI 16 Channel Async Line Multiplexor
1 450A Sun Microsystems second Ethernet controller (first

```
          controller included on CPU board)

Peripherals:
     2 Sun 626A Fujitsu M2361A Disk Drives     550 MB each
       CDC 9720-850 Disk Drive                 675 MB each
          1775 MB (1.7 GB) Total disk storage
     1 Sun 675A Fujitsu 6250/1600 BPI 1/2 inch tape drive

7 Sun Microsystems 3/60M Monochrome, diskless Workstations, each
       workstation configured with 12MB main memory
```

The 7 diskless Sun 3/60's and the Sun 3/280 are attached to a thin ethernet dedicated to providing file service for the diskless 3/60's. The second ethernet interface on the 3/280 is connected to the IntelliGenetics backbone ethernet. All network traffic, not destined for the diskless 3/60's, is routed via this second ethernet interface, to leave the other network free for file service to the 3/60's. A Cisco Systems Gateway controls the routing of network traffic between the BIONET diskless Sun network, two other local ethernets at Intelligenetics, the Arpanet, and the soon to be established BarrNet connection. BIONET users of the Sun system have electronic mail, telnet remote login, and ftp file transfer access to the rest of the Internet. These services are also available between the Suns and the Dec-20, until the phaseout of the Dec-20 is complete.

The Sun 3/280 and the 7 Sun 3/60's are accessible via the X.25 Public Data Networks, Telenet and CompuServ (see description of X.25 networks above). This connection is implemented by connecting serial ports on our X.25 Host Pads (described above) to ports on an existing Cisco Systems Terminal Server. A terminal server is a device which allows users of serial terminal line ports to establish login connections over the ethernet to hosts on the network. Using this mechanism, we are able to distribute BIONET users across the 3/280 and 7 3/60's (plus any additional incremental resources that might be added in the future) in a way which is transparent to the users. We are working on a load balancing system, which will ensure that each time a user connects to BIONET, he/she will be connected to the most lightly loaded computer resource available on our network.

A diagram of the entire BIONET computer system and network is included in *Appendix VII.*

Public Data Network Connection.

BIONET is accessed principally over the Telenet[1] and Compuserve Public Data Networks (PDN). An X.25 PAD (packet assembler/disassembler) is located on-site for each PDN. This is known as the Host PAD, or HPAD. It provides individual terminal ports which are cross-connected to those on the DEC-2065. The Telenet trunk line operates at 9600 baud synchronously, and the PAD converts this into up to 16 asynchronous ports whose speed is typically 1200 baud. A handshaking protocol is employed to smooth over bursts of data during the multiplexing.

Connection to multiple Public Data Networks increases geographic accessibility, since areas which are served poorly by one of the PDN's are often served well by the other. Reliability is also improved by providing alternate access when service is poor or unavailable on one of the networks. Our connection to both PDN's provided incentive to each PDN to offer us lower rates.

[1] The Telenet Public Data Network is operated by U.S. Sprint.

ARPANET

BIONET maintains a connection to the Arpanet which is arranged through a DARPA-funded project with IntelliCorp. In exchange for our assistance with the mechanics of the connection to ARPANET, BIONET is able to make use of this connection for communications, especially electronic mail and file transfer. Since there are mail gateways from the ARPANET to many other communications networks, this connection greatly expands BIONET's reach-- including networks such as BITNET, EARN and CSNET in addition to the DoD Internet.

The BIONET DEC-2065 is connected to a local ethernet at IntelliGenetics via its NI20 ethernet interface. A gateway connects the local ethernet to the Arpanet gateway at IntelliCorp via two 19.2KB leased lines.

NSFNET

In order to increase the reliability and speed of access to the DARPA/NSF Internet, BIONET has plans to share a T1 (1.54 Megabits/Sec) link to Barrnet (Bay Area Regional Research Network). Barrnet is the regional NSF sponsored network for the San Francisco Bay area. We expect our Barrnet connection to be operational by early January of 1989. This takes on additional importnce because in recent months DARPA has taken some initial steps in the direction of scaling down the Arpanet, and there has been talk about a complete phaseout at some unspecified time in the near future.

3.1.5.3 Summary Statistics on Machine Use

Originally BIONET had access to 50% of the cpu of DEC 2065 computer and the rest was used by IntelliGenetics and its parent company IntelliCorp. With the availability of other computer resources at IntelliGenetics this situation has now changed so that BIONET's share of the computer has risen to 90%.

The cpu cycles of the DEC-2065 computer are allocated to the user community, including BIONET, by the system's class scheduler. This scheduler is given the percentage of the machine to allocate to each class of users. Any cycles not consumed by a given class ("windfall") are available to the rest of the user community. This method was chosen so that cpu cycles not consumed by one segment of the community could be used by other segments if needed, i.e., no cpu cycles are wasted if someone needs them.

Because the class scheduler also adds to the system overhead, the number of user categories has been reduced to two. This reduces system overhead and frees up computing resources. Currently the batch queue and all users except IntelliGenetics commercial time sharing customers are in class 0 with a 90% allocation of the machine. Commercial time sharing users are in the second class and have a 10% allocation of the machine.

The actual use of the machine by the BIONET community has been on average each month 88%, substantially greater than the 50% of the total cpu cycles originally allocated for BIONET. As an example, the percentage use of the machine for the month of November, 1988 is shown in Figure III-1.

The data for BIONET's percentage of system use are plotted in histogram form in Figure III-2. This figure demonstrates that BIONET has utilized well over 50% of the total cpu cycles used on the 2065, and routinely consumes over 85% of the total cpu cycles used on the system. Last year this statistic averaged about 80%.

In the following series of tables and figures, we provide further details on the actual use of the system by the BIONET community. Looking first at use of the system in prime time (8 AM - 8 PM, M-F, PST), data for cpu time and connect hours for the indicated segments of the community are given in Tables III-3 and III-4 by month, and totals. The cpu data in Table III-3 is also plotted in histogram form in Figure III-3. BIONET prime time cpu usage is up over 45% compared to the year before!

The main conclusion derived from these data is the BIONET resource is being heavily utilized. This clearly demonstrates the demand for the BIONET service and explains why we approached Sun Microsystems to obtain additional equipment.

The total number of connect hours, prime time (Table III-4), for the category BIONET Users is up over last year by 37%. This is due both to continued growth in the user community and because two batch queues were opened during the year to run jobs around the clock on the DEC. Initially these queues were used very heavily, but the advent of the FASTA-MAIL program which utilizes the Sun 3/280 computer for database searches has almost eliminated the use of the DEC batch facility for routine database searches.

The data for non-prime time (weekends and 8 PM - 8 AM M-F) are shown in Tables III-5 and III-6, and the data on cpu time are plotted in histogram form in Figure III-4. Non-prime time cpu usage by BIONET has increased by 82% over the same figure for last year! Non-prime time cpu usage remains higher than prime time because of the courtesy of users in scheduling their DEC batch jobs to run during the evening hours. In both the prime and non-prime time periods BIONET staff cpu usage has declined compared to last year as the staff has moved their work to the Sun computers. Accounting software is not yet available for these systems.

The data for total use of the Resource by BIONET are presented in Tables III-7 and III-8 and the total cpu time is summarized in Figure III-5. Overall cpu usage is up by 65% and total connect hours have increased by 36% compared to last year.

There can be no doubt when viewing these statistics that the BIONET Resource is valued and heavily utilized by the scientific community. Impressive growth in the user community (over 200 new labs this year) and the statistics on machine use indicate that BIONET may be the most heavily utilized resource funded by the NIH.

Summary data for use of our telecommunications network are presented in Table III-9 and Figure III-6 by month for the past 12 months' use of the Telenet and Compuserve networks. Total connect hours increased by 85% over last year! Note that network "connect hours" represent actual time spent using Telenet or Compuserve while "connect hours" in previous figures include batch jobs, staff usage, and system overhead. These latter factors are accounted for as time "connected" to the computer but do not represent connections via Public Data Networks. Local use of the BIONET direct-dial access lines are also not included in Table III-9 and Figure III-6.